MUNI FI

Document Visual Question Answering with CIVQA

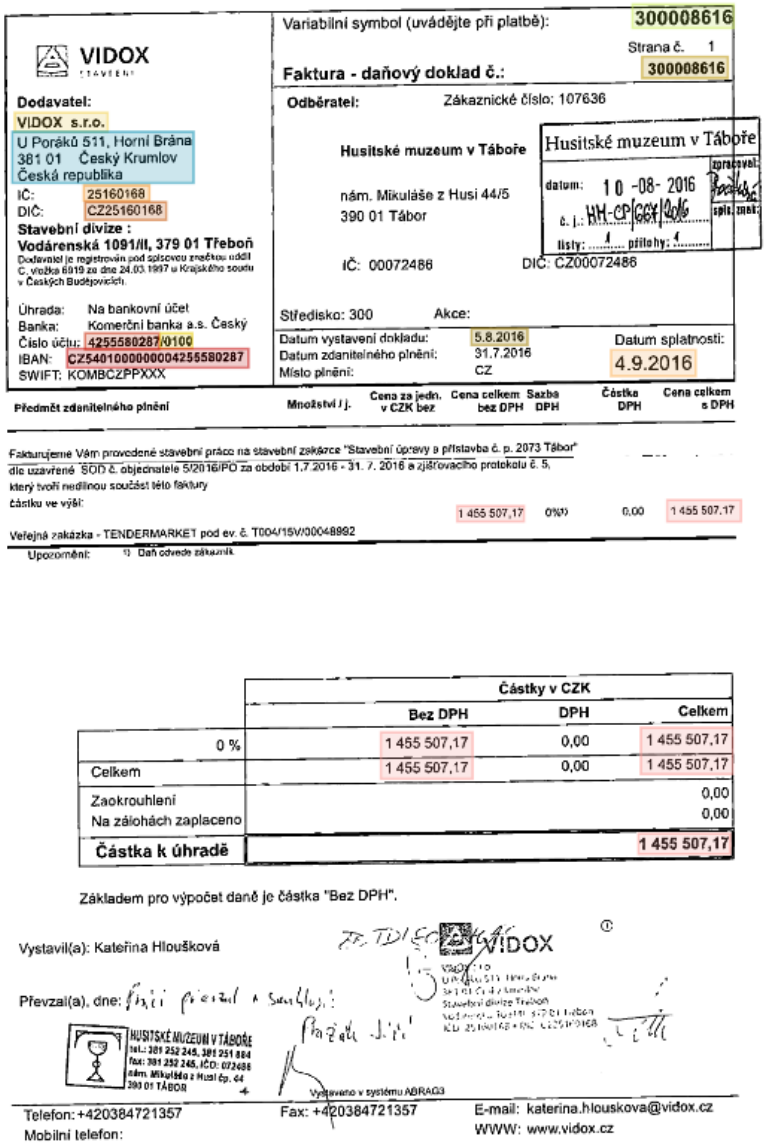
Czech Invoice Visual Question Answering Dataset

Šárka Ščavnická, Michal Štefánik, Petr Sojka

Faculty of Informatics, Masaryk University

December 8, 2023

Motivation



VIDOX

Variabilní symbol (uvádějte při platbě): 300008616

Strana č. 1

Faktura - daňový doklad č.: 300008616

Dodavatel:
VIDOX s.r.o.
U Poráků 511, Horní Brána
381 01 Český Krumlov
Česká republika
IČ: 25160168
DIČ: CZ25160168
Stavební divize :
Vodárenská 1091/II, 379 01 Třeboň
Dodavatel je registrován pod spisovou značkou oddíl C, vložka 6919 za dne 24.03.1997 u Krajského soudu v Českých Budějovicích.

Úhrada: Na bankovní účet
Banka: Komerční banka a.s. Český
Číslo účtu: 4255580287/0100
IBAN: CZ5401000000004255580287
SWIFT: KOMBCZPPXXX

Odběratel: Zákaznické číslo: 107636

Husitské muzeum v Táboře
nám. Mikuláše z Husi 44/5
390 01 Tábor

Husitské muzeum v Táboře
datum: 10 -08- 2016
č. j.: HH-CP/GEX/2016
listy: 1 přílohy: 1
zpracoval:
spis. znak:

IČ: 00072486 DIČ: CZ00072486

Středisko: 300 Akce:

Datum vystavení dokladu: 5.8.2016
Datum zdanitelného plnění: 31.7.2016
Místo plnění: CZ

Datum splatnosti: 4.9.2016

Předmět zdanitelného plnění	Množství / j.	Cena za jedn. v CZK bez	Cena celkem bez DPH	Sazba DPH	Částka DPH	Cena celkem s DPH
Fakturujeme Vám provedené stavební práce na stavební zakázce "Stavební úpravy a přístavba č. p. 2073 Tábor" dle uzavřené SOD č. objednatele 5/2016/PO za období 1.7.2016 - 31. 7. 2016 a zjišťovacího protokolu č. 5, který tvoří nedílnou součást této faktury						
částku ve výši:			1 455 507,17	0%1)	0,00	1 455 507,17

Veřejná zakázka - TENDERMARKET pod ev. č. T004/15V/00048992

Upozornění: 1) Daň odvede zákazník

	Částky v CZK		
	Bez DPH	DPH	Celkem
0 %	1 455 507,17	0,00	1 455 507,17
Celkem	1 455 507,17	0,00	1 455 507,17
Zaokrouhlení			0,00
Na zálohách zaplaceno			0,00
Částka k úhradě			1 455 507,17

Základem pro výpočet daně je částka "Bez DPH".

Vystavil(a): Kateřina Hloušková

Převzal(a), dne:

VIDOX

HUSITSKÉ MUZEUM V TÁBOŘE
tel.: 381 252 245, 381 251 884
fax: 381 252 245, IČO: 072486
nám. Mikuláše z Husi čp. 44
390 01 TÁBOR

Vystaveno v systému ABRAS3

Telefon: +420384721357 Fax: +420384721357 E-mail: katerina.hlouskova@vidox.cz
Mobilní telefon: WWW: www.vidox.cz

Visually rich documents and Document visual question-answering

- VRD contains such documents whose semantic structure is not determined only by the text but also by the layout and visual elements of the documents
- DVQA seeks to obtain knowledge from the documents' visual and textual parts to answer questions
- The asked questions may relate to different parts of the VRDs
  - text
  - inserted images
  - tables
  - forms

Used models

- LayoutLMv2
- LayoutXLM

  Chinese, Japanese, Spanish, French, Italian, German, and Portuguese
- LayoutLMv3
- Impira LayoutLM Invoices

  fine-tuned on the SQuAD and DocVQA datasets plus proprietary dataset of invoices
- Impira LayoutLM Document QA

  fine-tuned on the SQuAD and DocVQA datasets

CIVQA dataset

id string · *lengths* 9 9	words sequence	answers string · *lengths* 1 72	bboxes sequence	answers_bboxes sequence	questions string · *lengths* 4 35	image string · *lengths* 36 36
420000000	["12008626", "FAKTÚRA", "(FV)", "Strana", "1",…	Rosinská cesta 13 010 08 Žilina	[[78.69774919614147,…	[[15.434083601286174…	Jaká je adresa dodavatele?	f8f55985f6a82596baa43a72543fa4e4
420000001	["12008626", "FAKTÚRA", "(FV)", "Strana", "1",…	Rosinská cesta 13 010 08 Žilina	[[78.69774919614147,…	[[15.434083601286174…	Kde sídli dodavatel	f8f55985f6a82596baa43a72543fa4e4

Figure: CIVQA pre-encoded dataset

input_ids sequence	bbox array 2D	attention_mask sequence	image array 3D	start_positions int64 5 510	end_positions int64 7 510	questions string · *lengths* 4 35	answers string · *lengths* 1 72
[0, 4422,…	[[0, 0, 0, …	[1, 1, 1, 1, 1, 1, 1, 1…	[[[140, 145, 147, 149, 151, 153, 156, 156, 157, 159, 160, 163, 163, 161, 166, 169, 171, 170, 171,…	91	100	Jaká je adresa dodavatele?	Rosinská cesta 13 0 Žilina
[0, 119950,…	[[0, 0, 0, …	[1, 1, 1, 1, 1, 1, 1, 1…	[[[140, 145, 147, 149, 151, 153, 156, 156, 157, 159, 160, 163, 163, 161, 166, 169, 171, 170, 171,…	88	97	Kde sídli dodavatel	Rosinská cesta 13 0 Žilina

Figure: CIVQA encoded dataset

Entity	Numeric	Textual	Pattern	Shape
Invoice number	X			
Variable symbol	X			
Specific symbol	X			
Constant symbol	X			
Bank code	X			X
Account number	X			X
ICO	X			X
Total amount	X			
Invoice date	X			X
Due date	X			X
Name of supplier		X		
IBAN	X	X		X
DIC	X	X		X
QR code			X	X
Supplier's address		X		

Table: CIVQA dataset's entities' categories

Tesseract and EasyOCR CIVQA dataset

- Tesseract OCR
  - was developed at HP Research between 1984 and 1994
  - Open-source project since 2005
  - Can recognise more than 100 different languages, including Czech
- EasyOCR
  - Python framework created by Jaded AI
  - Can recognise just over eighty languages, including Czech
- Each type of these dataset has two different versions
  - Readable by human
  - Ready to use

Tesseract OCR vs EasyOCR

Table: CIVQA results: comparison of Tesseract and EasyOCR frameworks by Precision, Recall, and F1 score.

	Tesseract			EasyOCR		
Model	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**
LayoutXLM	**0.7422**	**0.7117**	**0.7079**	0.6636	0.6633	0.6455
LayoutLMv2	0.6917	0.6750	0.6634	0.6323	0.6129	0.6011
LayoutLMv3	0.6989	0.6382	0.6410	0.6370	0.6164	0.6065
Impira QA	0.6773	0.6291	0.6313	0.6373	0.6015	0.5984
Impira Invoice	0.6948	0.6440	0.6434	0.6345	0.6019	0.5962

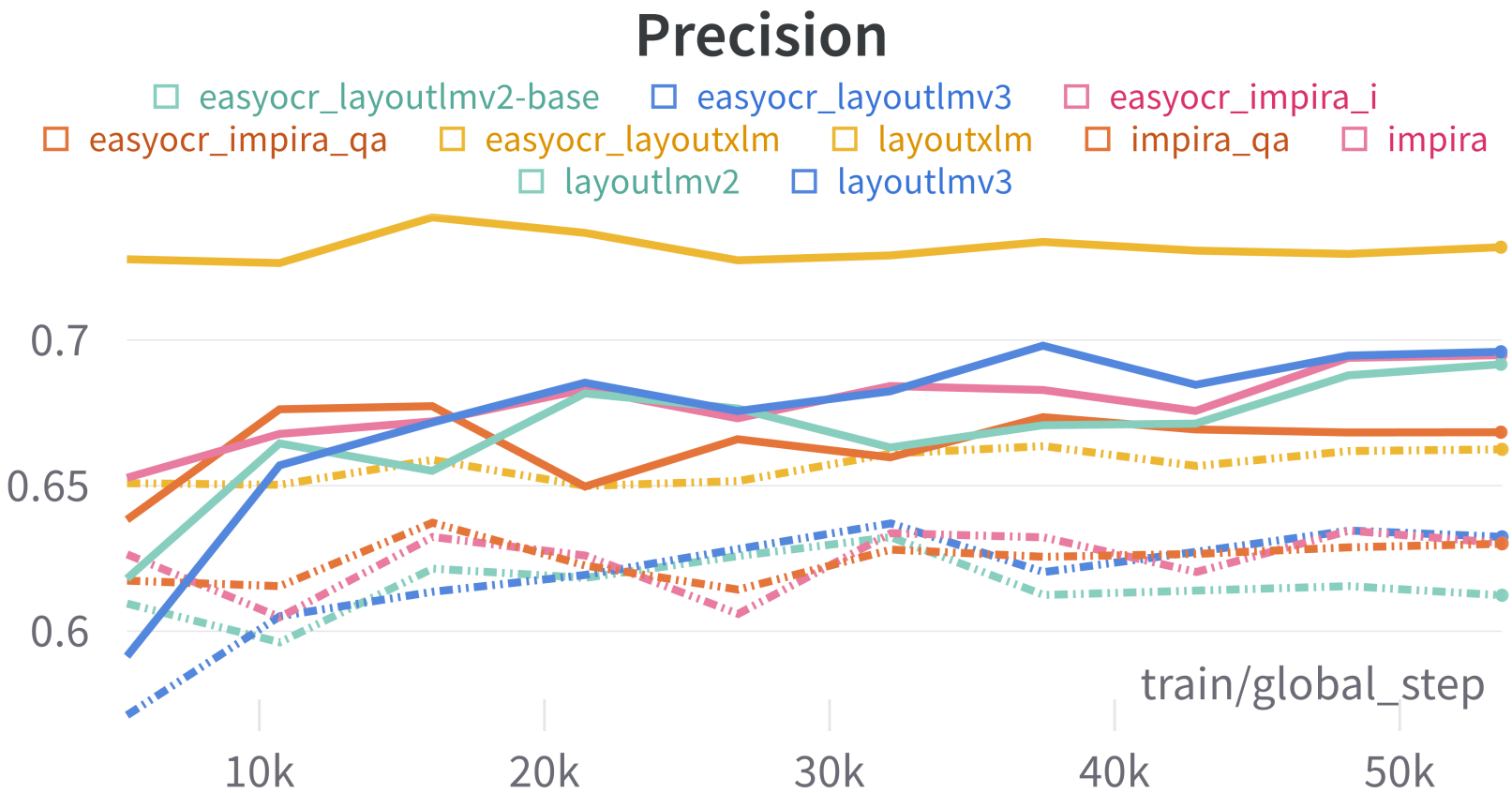
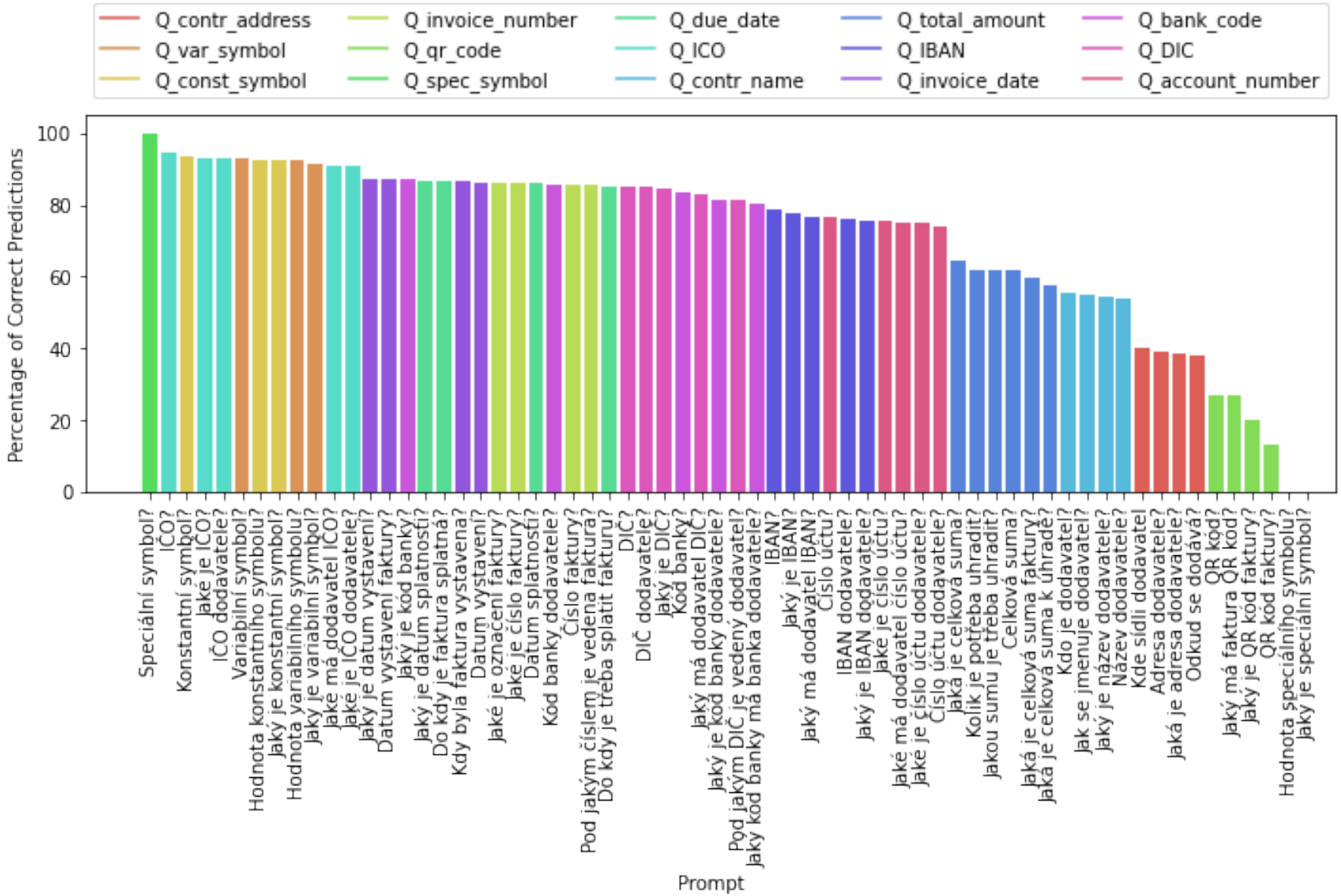


Figure: The precision of the models in the first experiment.


Q_contr_address
Q_var_symbol
Q_const_symbol
Q_invoice_number
Q_qr_code
Q_spec_symbol
Q_due_date
Q_ICO
Q_contr_name
Q_total_amount
Q_IBAN
Q_invoice_date
Q_bank_code
Q_DIC
Q_account_number
Percentage of Correct Predictions
0
20
40
60
80
100
Speciální symbol?
IČO?
Konstantní symbol?
Jaké je IČO?
IČO dodavatele?
Variabilní symbol?
Hodnota konstantního symbolu?
Jaký je konstantní symbol?
Hodnota variabilního symbolu?
Jaký je variabilní symbol?
Jaké má dodavatel IČO?
Jaké je IČO dodavatele?
Jaký je datum vystavení?
Datum vystavení faktury?
Jaký je kód banky?
Jaký je datum splatnosti?
Do kdy je faktura splatná?
Kdy byla faktura vystavena?
Datum vystavení?
Jaké je označení faktury?
Jaké je číslo faktury?
Datum splatnosti?
Kód banky dodavatele?
Číslo faktury?
Pod jakým číslem je vedena faktura?
Do kdy je třeba splatit fakturu?
DIČ?
DIČ dodavatele?
Jaký je DIČ?
Kód banky?
Jaký má dodavatel DIČ?
Jaký je kód banky dodavatele?
Pod jakým DIČ je vedený dodavatel?
Jaký kód banky má banka dodavatele?
IBAN?
Jaký je IBAN?
Jaký má dodavatel IBAN?
Číslo účtu?
IBAN dodavatele?
Jaký je IBAN dodavatele?
Jaké je číslo účtu?
Jaké má dodavatel číslo účtu?
Jaké je číslo účtu dodavatele?
Číslo účtu dodavatele?
Jaká je celková suma?
Kolik je potřeba uhradit?
Jakou sumu je třeba uhradit?
Celková suma?
Jaká je celková suma faktury?
Jaká je celková suma k úhradě?
Kdo je dodavatel?
Jak se jmenuje dodavatel?
Jaký je název dodavatele?
Název dodavatele?
Kde sídlí dodavatel
Adresa dodavatele?
Jaká je adresa dodavatele?
Odkud se dodává?
QR kód?
Jaký má faktura QR kód?
Jaký je QR kód faktury?
QR kód faktury?
Hodnota speciálního symbolu?
Jaký je speciální symbol?
Prompt

Two lines answers

	prompt	prediction	true
1	Jaká je adresa dodavatele?	Rosinská cesta 13 010 08 Žilina	Rosinská cesta 13 010 08 Žilina

Figure: The correct answer is on one line.

	prompt	prediction	true
99	Jaká je adresa dodavatele?	Michalovská 62 Variabilný symbol: 1858420000 073 01 Sobrance	073 01 Sobrance

Figure: The correct answer is on multiple lines, so it was split.

CIVQA and unseen types of questions

In this set of experiments, our focus was on developing a practical and robust solution for unseen entities.

- **Invoice number**

  A numerical entity without a fixed shape.
- **ICO**

  A numerical entity with given shape.
- **Supplier's address**

  Textual and numerical entity without a fixed shape.
- **IBAN**

  Textual and numerical entity with a fixed shape.
- **Due date**

  A numerical entity with given shape.

Baseline on unseen entities

Model	Precision	Recall	F1 score
LayoutXLM	0	0	0
LayoutLMv2	0	0	0
LayoutLMv3	0	0	0
Impira QA	0	0	0
Impira Invoice	0	0	0

Table: CIVQA Tesseract OCR results on unknow entities

Baseline on unseen entities

Table: CIVQA results: comparison of models when handling unknown entities

	Known data			DocVQA + Known data		
Model	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**
LayoutXLM	0.1920	0.0413	0.0582	**0.3731**	**0.2163**	**0.2465**
LayoutLMv2	0.0343	0.0270	0.0261	0.0665	0.0334	0.0279
LayoutLMv3	0.1022	0.0341	0.0456	0.1504	0.0455	0.0611
Impira QA	0.1512	0.0455	0.0652	0.2326	0.0895	0.1148
Impira Invoice	0.1360	0.0530	0.0724	0.2226	0.0807	0.1063

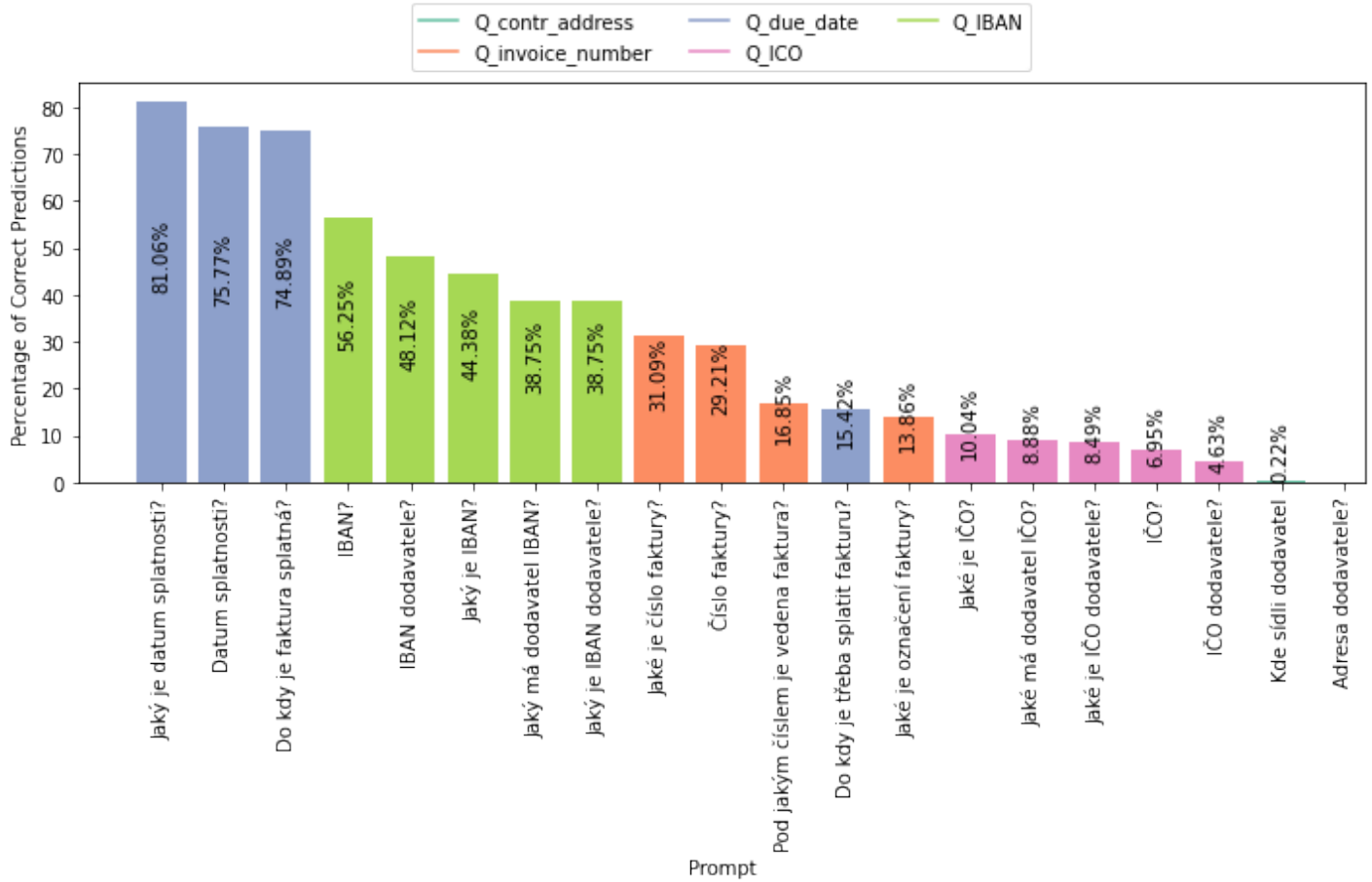


Figure: Validation dataset of CIVQA unknown entities: LayoutXLM model success rate by individual question percentage fine-tuned on DocVQA plus CIVQA known dataset.

Training with a subset of unknown data

Table: CIVQA results: Comparing results on baseline models, then models trained on a 5% subset of unknown entities and then models fine-tuned on the concatenation of known dataset with a subset of 5% unknown entities.

	Known data			5% of unknown			Known + 5% unknown		
Model	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**	**Prec**	**Recall**	**F1**
LayoutXLM	0.1920	0.0413	0.0582	0.7002	0.6594	0.6617	**0.7069**	**0.6693**	**0.6700**
LayoutLMv2	0.0343	0.0270	0.0261	0.5944	0.5154	0.5192	0.6223	0.5726	0.5755
LayoutLMv3	0.1022	0.0341	0.0456	0.5793	0.5125	0.5254	0.6344	0.5528	0.5631
Impira QA	0.1512	0.0455	0.0652	0.6186	0.5356	0.5466	0.6318	0.5487	0.5670
Impira Invoice	0.1360	0.0530	0.0724	0.5999	0.5255	0.5369	0.6353	0.5577	0.5681

Conclusion

- CIVQA dataset
- Numeric answers obtained better results than purely textual ones
- Entities with given structure perform betters
- LayoutXLM

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS